

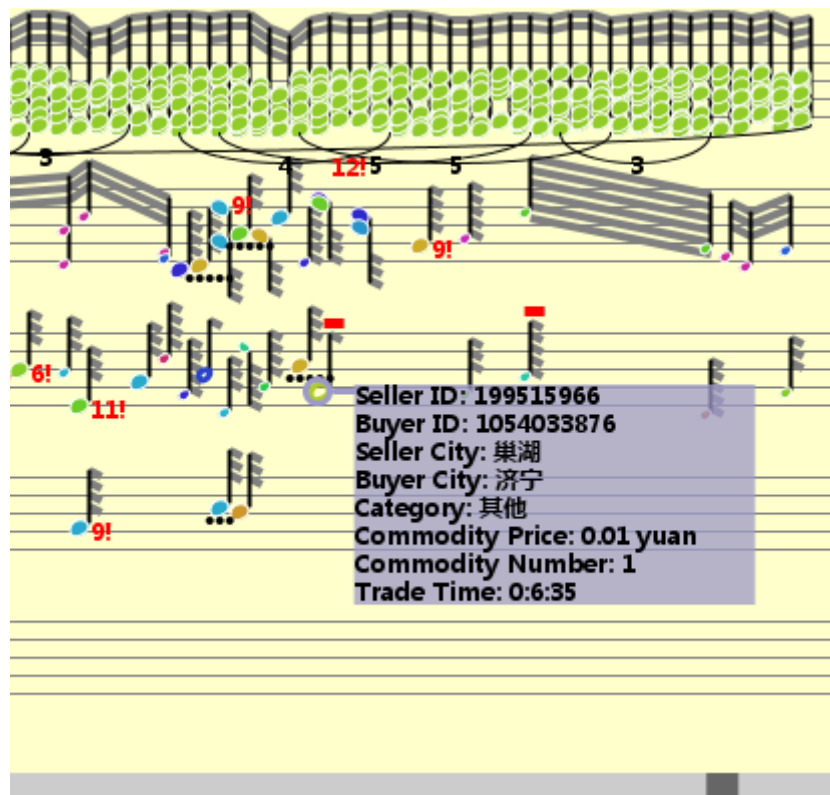
## 本周周报(10.22-10.28):

解聪

本周工作:

### 1. 淘宝交易地图

本周在上周的设计基础上添加了一些交互, 包括点击显示相关信息以及滚动显示等等。如下图所示;



本周重点研究了数据分析的问题。主要包括两个方面:

### 1. 从一些论文中寻找可以借鉴的方法:

在重新了解“A Symbolic Representation of Time Series, with Implications for Streaming Algorithms”的方法之后, 我感觉这篇文章的方法完全适合交易数据的特性 --- 大数据量, 高维度, 离散的数据。

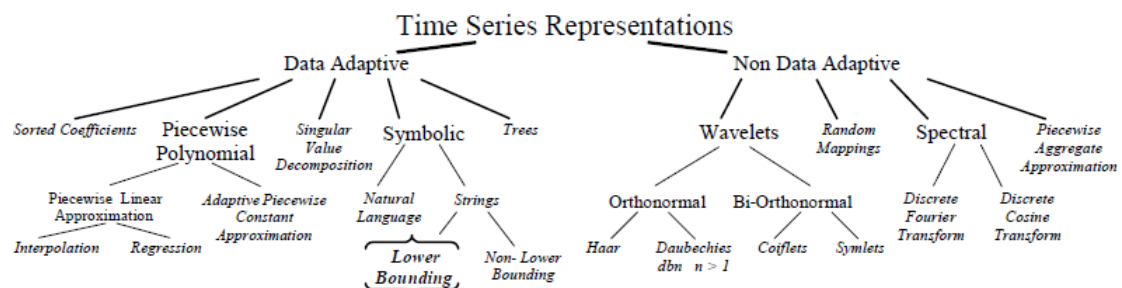


图 文中提出的时间序列的处理方法。分类十分详细。

文章中提出使用离散的字符串来代替时序数据,好处有:可以处理较大的数据量:可达10亿级;可以使用字符串处理的手段:如后缀树;将每种字符看成状态,可以使用马尔科夫模型或决策树等方法来处理,发现其中的异常数据。

作者引文中的两篇相关文章还详细介绍了如何对时序数据进行主题抽取,与先前的方法相比,其可以在线性时间内完成处理。这对交易数据的主题抽取很有帮助。联系这一系列论文,完全可以为我们的问题提出了一个很好的解决方案。先使用离散的交易数据采用字符串序列的方法,再提取序列中的一些特定主题。

现在的问题就是,如何由高维的交易数据构造转化为便于文中处理的字符串序列。这其中的难点又重新回到了如何定义异常性的问题。在前几周的工作中,我考虑了一些因素,但是也不太理想。

## **2. 向淘宝数据安全部门的人询问有关信息:**

联系到安全部门的一个同事玄骏,他们恰好有在做异常交易检测的工作。

首先初步了解了他们的方法:他们拥有丰富的数据维度,并采用了比较复杂的模型来进行判断刷信誉的交易。他们处理过程中,除了利用交易本身的属性外,还使用了更多交易外的信息,包括有通讯软件的对话信息,交易的 IP 之间的联系,两个 ID 之间的关系等等。但是就我们现有的数据来说是无法满足这样的要求。初步了解过程中,尚未了解到他们的复杂的模型是怎么够早的,但是核心数据挖掘算法使用了决策树。

在初步了解了我们的工作后,他觉得我们的想法很有创意。但是他觉得数据分析这一步十分复杂,开发周期达到几个月。从他们的角度以及成本节约的角度,他有建议我直接使用可以搜集到的分析好的数据。目前暂时预约了下周和他面谈,重点了解他们的方法在我们现有的数据上是否有可以借鉴的地方和创新点。询问一些可以使用的方法以及其好处。

## **2. DataV 组件**

绘制雷达图。

## **下周工作:**

### **1. 淘宝交易地图**

1. 继续探索使用何种数据分析方式。
2. 与玄骏了解异常交易的定义,处理大数据中寻找异常的一些方法,以及这些方法的特点。
3. 完善现有可视化的交互界面。

### **2. DataV 组件的开发**

完善雷达图。